The Ohio State Technology Law Journal

ANALYTICAL CHALLENGES IN MODERN TAX ADMINISTRATION

A BRIEF HISTORY OF ANALYTICS AT THE IRS

JEFF BUTLER*

CONTENTS

* Internal Revenue Service, Research, Applied Analytics, and Statistics. The author would like to thank Bryan Choi and Bill Roberts for their careful reading of the manuscript and thoughtful comments. Additional thanks to Mike Dunn, Ron Hodge, Rahul Tikekar, and Anne Parker for vigorous and valuable discussions which helped guide the exposition.

I. INTRODUCTION

The IRS has a long history of using data for decision making to meet its statutory mandates and improve business operations. Throughout its more than 100-year existence, for example, the IRS Statistics of Income Division has been using stratified and complex sampling techniques to produce tax statistics for the IRS, Congress, and the public. The use of ordinary least squares regression to forecast tax returns volumes was first applied more than 50 years ago by the Research Division to support planning and budgeting functions. In the 1970s, linear discriminant analysis was introduced as a method to identify non-compliant tax returns for examination, an effort that resulted in the Discriminant Index Function (DIF) program. While only one of several techniques used for audit selection today, DIF was considered revolutionary at the time.

In an early effort to reach beyond traditional statistics, the IRS established an artificial intelligence lab in the mid-1980s under the aegis of its research division. Primary focus was on symbolic reasoning, natural language processing, and simple feed-forward (e.g., Hopfield) neural networks.[1] Solutions to several use cases were field-tested on a limited basis in the 1990s, including a knowledge base keyword search for a district office call center, an expert system to detect non-compliance on Earned Income Tax Credit (EITC) returns, and a web-based, automated tax law assistant that had the primitive features of a modern question-answering system. While it was never put into operation, the Automated Issue Identification System (AIIS) was an expert system designed to detect anomalies on a composite of Form 1040 line items, becoming the first rules-based alternative to DIF. Considering the rapid progress of artificial intelligence and its subfields today, most recently deep learning, the idea of an AI lab in the IRS nearly 35 years ago appears somewhat visionary in hindsight.[2]

---

[1] CLIFFORD LAU, NEURAL NETWORKS: THEORETICAL FOUNDATIONS AND ANALYSIS 27 (1992); EDGAR SÁNCHEZ-SINENCIO, ARTIFICIAL NEURAL NETWORKS: PARADIGMS, APPLICATIONS, AND HARDWARE IMPLEMENTATIONS (Clifford Lau ed., IEEE Press 1992).

[2] The AI lab was dissolved in the mid-1990s.

Outside of the AI lab, a number of other analytical initiatives also emerged. Linear programming was used to create a post-of-duty model to optimize geographic locations for walk-in sites, employment offices, and other facilities. A rules-based algorithm was developed to identify duplicate dependent SSNs on separate tax returns claiming EITC. This model would evolve into the Dependent Database (DDb), the primary IRS system today for detecting refundable credit fraud. A new statistical framework based on decision trees was introduced for classification of payment risk and case routing in collection workload. This was an effort that ran a Kaggle-like competition of techniques that, in addition to recursive partitioning, also included k-nearest neighbor discrimination, kernel discriminant analysis, neural networks, and a family of standard linear models. It was also during the 1990s that the Compliance Data Warehouse was created, which remains the largest analytical computing environment in the IRS to this day.

In the 2000s, graph theory was used for the first time to analyze flow-through income reported on the Schedule K-1 from Partnerships, Sub-Chapter S corporations, and Trusts.[3] The result was a breakthrough tool, called the YK-1 Link Analysis System, that gave IRS revenue agents the ability to quickly visualize complex, connected entities. This initiative also involved the first known use in the IRS of support vector machines for classification. New decision tree models were developed to identify tax return fraud shortly after the IRS initiated electronic filing, an effort that eventually resulted in the Electronic Fraud Detection System (EFDS), forerunner to the Return Review Program (RRP) in use today. Micro-simulation models of taxpayer burden were developed to analyze the distributional impact of burden for different subgroups based on changes in tax policy or administrative procedures. Agent-based models were created for the first time to study the self-organizing dynamics of non-compliance behavior. Other noteworthy projects during this period that used novel techniques included rule induction and Benford's Law, both of which were aimed at detecting reporting noncompliance on tax returns.[4]

---

[3] Dave Debarr & Zach Eyler-Walker, *Closing the Gap: Automated Screening of Tax Returns to Identify Egregious Tax Shelters*, 8 SIGKDD EXPLORATIONS 11, 11 (2006).
[4] MARK J. NIGRINI, BENFORD'S LAW: APPLICATIONS IN FORENSIC ACCOUNTING, AUDITING, AND FRAUD DETECTION 291 (2012).

Following a deep tradition of leveraging domain expertise to encode knowledge through rules, the period from 2010-2015 produced a number of new rule-based models for identity theft, refund fraud, preparer non-compliance, and federal tax deposit payments. Legislation that created information reporting on the Form 1099-K drove the development of a large, multi-stage clustering model for anomaly detection and case selection. This was one of the first significant attempts to use unsupervised learning for such a purpose. It was also during this period that the first native graph database was used to identify entity fabrication and pyramiding schemes.

II. ANALYTICAL NEEDS OF TAX ADMINISTRATION TODAY

Tax administration faces a wide range of analytical issues: identifying hidden relationships in complex business structures and multi-party transactions; finding identity theft in multiple low-dimensional subspaces; detecting anomalies in sparse, unlabeled tax return data; classifying multi-class taxpayer issues through case notes, administrative decisions, and correspondence; and the management of massive amounts of data from distributed, heterogeneous, and multi-structured formats. These and other problems are directly linked to strategic mission objectives of the IRS. In recent years, there has been renewed interest among researchers to look for new and non-traditional models, enabled by increasing amounts of data and computing power, to address important questions that were once considered off-limits, including:

- How do we model tax return anomalies in high-dimensional spaces, often sparse, to detect a class of non-compliance when there are no prior class labels?

- When does a subgraph of a business structure and its related entities fail to represent a typical, unbiased sample of the underlying population of similar entities?

- How should we further isolate and classify those substructures into distinct subclasses?

- When is a posterior probability of payment in a branching sequence of transactions a signal rather than a random event or con-founding factor?

- How do we quantify interrelated factors that drive service interactions and predict treatment options with the best likelihood of resolution?

- How do we make predictions from the wide range of temporal disturbances observed in tax data, e.g., changes in preparer communities, dependent child status, ownership structures, and so on?

- How do we determine whether those disturbances are transient or permanent?

While these questions are not intended to represent a set of current strategic priorities, they are examples of important and difficult problems that will require non-trivial methodologies beyond classical statistics and the development of new methods that can improve on existing foundations. Addressing these types of questions will require a cross-cutting approach that brings to bear new data and tools for advanced analytics and AI. It will demand changes necessary to overcome obstacles, sometimes institutional, in order to realize the extraordinary opportunities at hand.

III. CURRENT STATE OF ANALYTICS AT THE IRS

In the past several years, the IRS has made progress through partnerships with industry and academia to incubate and test a range of new analytical approaches, such as those with a renewed emphasis on AI subfields like machine learning, natural language processing, knowledge representation, and evolutionary systems. Key factors that have influenced this change include:

- Increasing accessibility to a vast amount of new data from large-scale collections of heterogenous, multi-structured data sources.

- Advances in cutting-edge tools for analytics that support a diverse set of use patterns, some of which are computationally complex.

- Better technology and computing architectures that reflect the needs of modern analytics including: large-scale warehouses of historical records; highly distributed parallel systems; GPU-based computing frameworks; and large memory systems for in-memory computing of complex data structures.

- Extended talent pools through partnerships with industry and academia in advanced analytics, data science, and artificial intelligence.

External influences have also motivated the need for new methodological approaches, including research and agency recommendations on applications of big data, data science, and artificial intelligence from the National Academy of Sciences[5] and National Science and Technology Council[6]; new legislation requiring agencies to improve data and analytical governance[7]; and a 2019 White House Executive Order outlining steps to maintain U.S. leadership in AI.[8]

While there is still a substantial amount of observational analysis based on simple statistical methods, sometimes improperly used, an emerging set of diverse and sophisticated methods is being established:

---

[5] *See, e.g.,* NAT'L RES. COUNCIL, FRONTIERS IN MASSIVE DATA ANALYSIS (The Nat'l Acad. Press ed., 2013).
[6] NAT'L SCI. & TECH. COUNCIL COMM. ON TECH., EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (Oct. 12, 2016).
[7] Foundations for Evidence-Based Policymaking Act of 2018, H.R. 4174, 115th CONG. (2018).
[8] Exec. Order No. 13859, 84 Fed. Reg. 31 (Feb. 11, 2019).

Machine Learning and Deep Learning. Decision trees and random forests have been used to improve the detection of identity theft and refund fraud; discriminant analysis and naïve Bayes are among the many methods used to detect non-compliance on tax returns; logistic regression has been used for a variety of classification problems, including case routing in Collection and attrition models for human capital planning; and survival analysis has been evaluated for predicting payments in accounts receivable. Linear and mixed integer programming are used for resource allocation decisions, case ranking, and scenario building in Examination. Clustering and principal components analysis have been applied in a number of areas for data summarization, feature engineering, and anomaly detection. Convolutional neural networks are being used to identify and extract specific structures from IRS Appeals documents to support Examination planning.

Recommender Systems. Field testing of recommender systems was recently initiated in Examination for anomaly detection and case ordering, and interest exists in its application to journey maps in customer service. Early testing suggests that blended collaborative filtering models based on matrix factorization[9] can improve root mean square errors from existing supervised classification models, but without the cost of class labels needed for supervised learning.

Natural Language Processing. Text classification is increasingly being used to analyze tax forms, attachments, internal decision documents, case notes, and third party information. Entity extraction has been applied to Bank Secrecy Act reports, Department of Justice documents, and reportable transactions. Topic modeling is used to analyze corporate change of accounting statements and IRS Appeals decisions. Fuzzy matching has been used as first-stage input to models of tax preparer networks. Word embeddings have recently been introduced to model case notes from customer service interactions.

---

[9] CHARU C. AGGARWAL, RECOMMENDER SYSTEMS: THE TEXTBOOK 91 (2016); *see generally* William J. J. Roberts, *Application of a Gaussian, Missing-Data Model to Product Recommendation*, 17 IEEE SIGNAL PROCESSING LETTERS 509 (2010).

Language translation algorithms are being tested on treaty partner data, and interest exists in the systematic development of question-answering systems for web-based interactions.

Graph Mining. In the past few years, several graph databases have been created: a 100-million node graph of business entities originally designed to analyze entity fabrication and pyramiding; a graph to analyze temporal irregularities of preparer networks through community detection; and a graph of excise activity to better understand asset flows and multi-party transactions. Work has started on a graph of U.S. multinationals to analyze changes in corporate ownership, a problem that is prohibitively difficult with relational data structures. Graph mining techniques, including link analysis, graph matching, and anomaly detection,[10] are being evaluated in several areas. Graphs have also been introduced as policy networks[11] to proactively plan for the cascading impact of new tax legislation on policies, procedures, IT systems, and tax forms and publications.

Evolutionary Systems and Simulation. Several discrete simulation models have been developed over the years to better understand and predict events and outcomes for selected business processes related to non-filer, taxpayer assistance center, Whistleblower, and Affordable Care Act programs. Agent-Based Models have been used to study a range of taxpayer behaviors,[12] including tax evasion, and the IRS is making new investments in computational infrastructure, training, and partnerships for future applications of large-scale ABMs. Genetic algorithms have been developed to simulate self-organizing tax schemes based on observed patterns of shelters, off-shore activity, and other tax avoidance strategies.[13]

---

[10] Leman Akoglu, Hanghang Tong & Danai Koutra, *Graph-Based Anomaly Detection and Description: A Survey,* 29 DATA MINING & KNOWLEDGE DISCOVERY 626, 627 (2014).
[11] William D. Coleman, *Policy Networks*, in INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES (James D. Wright et al. eds., 2d. ed. 2015).
[12] AGENT-BASED MODELING OF TAX EVASION: THEORETICAL CONCEPTS AND COMPUTATIONAL SIMULATIONS xxvii (Sascha Hokamp et al. eds., 2018) [hereinafter AGENT-BASED MODELING OF TAX EVASION].
[13] Jacob Rosen et al., *Detecting Tax Evasion: A Co-Evolutionary Approach*, 24 J. ARTIFICIAL INTELLIGENCE & L. 149, 161 (2016); Geoffrey Warner et al., *Modeling Tax Evasion With Genetic Algorithms*, 16 ECON. GOVERNANCE 165, 165-66 (2014).

IV. DATA AND TECHNOLOGY NEEDED TO ADVANCE ANALYTICS

Progress has been made in the IRS over the past decade to collect and store an increasing amount of relevant data for research and analytics. Long historical records with over 20 years of data now exist for key sources including tax returns, customer accounts, information returns, and select compliance histories from case management systems. Many useful data sets for analytics and tax research have been produced by combining multiple, disparate data sources, including:

- 11-billion row, 20-year longitudinal panel created by IRS, Treasury, and academic partners to simplify the analysis of relationship-records for 1040 filers.

- 100-million node graph database of business entities to better understand business formation and complex relationships between connected entities.

- 1.2 billion row data set to analyze improper credits claimed on returns submitted by tax preparers.

While new insights and discovery are being made from these and other examples, collaborative development of new data sets is needed to address a range of open and important questions in tax administration. Examples include: an integrated history of customer service data from voice recordings, web transactions, account history, walk-ins, and write-in correspondence to improve classification of taxpayer service interactions; graph data structures for identity theft and refundable credit fraud; large-scale synthetic data for public use; domain-specific corpora needed to train NLP models; new data structures for analyzing increasing complexities of international taxation that incorporates third-party and treaty partner information; and baseline data that can be used for large-scale simulation of emergent tax schemes. Despite recent progress, the use of unstructured data remains particularly

limited. The IRS is awash in usable text from tax forms and attachments, treaty partners, John Doe Summons, case notes, customer service channels, criminal records, and procedural or administrative documents. More work is needed to remove barriers to access these data and combine them in multi-modal representations to create new opportunities for discovery.

A number of inherent challenges are associated with analyzing massively large data sets, and recent efforts have been made to invest in state-of-the-art technology needed to innovate and advance research and applications. Traditional data warehousing with columnar data stores has been augmented by large memory systems, GPU-based computing, distributed architectures, flash storage, and a faster network fabric. In the IRS research environment, open source software now dominates what was once a monopoly of commercially licensed tools. Design considerations continue to be shaped by a full range of use patterns in analytics and AI to generate the greatest benefits from the vast amount of available data.

While the IRS has evolved its research and analytical computing infrastructure, investment is needed to scale up and keep pace with the size and growth of data, and to maximize capabilities for cutting-edge tax administration research. In order to fully utilize state-of-the-art computing capabilities, greater training and workforce skills are necessary in non-relational data structures, distributed and parallel computing frameworks (e.g., CUDA, OpenCL, MPI, and OpenMP), and in open source languages and APIs designed for the kind of distributed, large-scale processing needed for massively large data sets or computational complexity.

V. SOME EMERGING CHALLENGES FOR THE COMING DECADE

While it is difficult to make predictions about the direction and priorities of research and analytics needed in the IRS in the long run, a small number of emerging and relatively important challenges can be discussed.

a. Large-scale Simulation of Events and Agents

Many problems in tax administration do not lend themselves to controlled experiments. Depending on the scale and complexity, some problems are simply too difficult or expensive for controlled experiments.; others may take too long or be considered too risky. For example, it would be too costly for the IRS to create multiple, parallel business processes to test the impact of changes to toll-free telephone service on say, half of all calls—which would involve tens of millions of taxpayers. A live field test of different scenarios for tax return examinations on a significant percent of the audit inventory would be considered too risky. Changing the rules for loss limitations and passive income for a small group of partnerships to measure network diffusion effects is legally impractical.

While the IRS has introduced new tools, computing architectures, and technology for simulation modeling, important challenges still remain. One is developing a unified framework for discrete event simulation (DES) that integrates system dynamics across business processes that are currently analyzed independently. In theory, enforcement and service activities are interdependent. How do we simulate connections between desired target states of both enforcement (e.g., utilization, revenue) and customer service (e.g., wait times, satisfaction)? How can we optimize a given state in Collection by changing input constraints for cases created in Examination? Methodologically, a unified design for these types of questions may also need to consider time and scale heterogeneities. Enforcement activities can take years, while service events might only take minutes.

Modeling the behavior of huge social networks through agent-based modeling (ABM) is another opportunity that is getting more attention.[14] Traditionally, simulation models of social systems at the macro-scale are designed to have complex, non-linear structure with fewer entities. At the micro-scale, they rely on simpler equations with

[14] AGENT-BASED MODELING OF TAX EVASION, *supra* note 12, at xvi.

local rules, mostly linear, with a larger number of entities.[15] In 2018, the IRS received tax returns from over 150 million individual taxpayers who are variously linked to dependents, third-parties, preparers, employers, and other entities. Even for experiments not involving multiple time periods, ABMs of this size present computational limitations on the ability to generate artificial networks with sufficient complexity to understand the full range of evolutionary behavior. Dynamic rule ecologies based on genetic algorithms, neural networks, or other self-generating mechanisms may only add to computational cost.[16] Designing an architecture to overcome these limitations would create new opportunities for future work.

b. Applications of Deep Learning

The main objective of machine learning is to identify and classify features from data without human involvement. While many machine learning algorithms share common optimization strategies with statistical models (e.g., L2-norm minimization), their objective emphasizes algorithms that accurately predict patterns regardless of inferences about the underlying probability generating mechanism. This essential difference has been a key driver for big data and data science paradigms.[17] Deep learning extends this emphasis through representational learning that further abstracts the use of the backpropagation algorithm through additional transformations of the input data.[18]

Deep learning has proven to be exceptionally successful for problems in image processing, speech recognition, and language translation.[19]

---

[15] See Eric Silverman & John Bryden, *From Artificial Societies to New Social Science Theory*, in 9 ADVANCES IN ARTIFICIAL LIFE 565 (2007).
[16] JOSHUA M. EPSTEIN & ROBERT AXTELL, GROWING ARTIFICIAL SOCIETIES: SOCIAL SCIENCE FROM THE BOTTOM UP 162 (1996).
[17] Leo Breiman, *Statistical Modeling: The Two Cultures,* 16 STAT. SCI. 199, 213 (2001); David Donoho, *50 Years of Data Science*, 26 J. COMPUTATIONAL & GRAPHICAL STAT. 745, 763 (2017).
[18] Li Deng, *A Tutorial Survey of Architectures, Algorithms, and Applications for Deep Learning*, 3 APSIPA TRANSACTIONS SIGNAL & INFO. PROCESSING 1, 12 (2014).
[19] *See generally* Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436 (2015).

While it may be too early to speculate on the full range of applications for tax administration, several use cases should be considered. Convolutional neural networks could be used to detect unusual patterns of toll-free calls by analyzing composite data based on call duration, number of daily calls, question categories, call flow point density, and other criteria.[20] Graph embeddings and graph convolutional networks, which belong to an emerging learning paradigm known as geometric deep learning,[21] have potential applicability for anomaly detection[22] in complex business structures (e.g., partnerships), category prediction in policy networks, recommendations in customer journey maps, and as an alternative to traditional graph classification.

Recurrent neural networks with long short-term memory are often used to construct a sequence of word embeddings based on the probability distribution of prior states to predict the next character in a word, or the next word in a sentence.[23] Applied to calls, correspondence, or narrative in tax forms, this technique could potentially be used for anomaly detection. For example, are large prediction errors in narratives of amended returns a good indicator of which ones are referred to Examination?

c. Synthetic Generation of Tax Data for the Public Domain

Giving scientific practitioners in the public domain greater access to administrative tax data for research would accelerate insights and

---

[20] Alae Chouiekh & EL Hassane Ibn EL Haj, *ConvNets for Fraud Detection Analysis*, 127 PROCEDIA COMPUTER SCI. 133, 134-38 (2018).

[21] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam & Pierre Vandergheynst, Geometric Deep Learning: Going Beyond Euclidean Data 1 (May 3, 2017) (unpublished manuscript), https://arxiv.org/pdf/1611.08097.pdf; Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang & Philip S. Yu, A Comprehensive Survey on Graph Neural Networks (Aug. 8, 2019) (unpublished manuscript), https://arxiv.org/pdf/1901.00596.pdf.

[22] Kaize Ding et al., *Deep Anomaly Detection on Attributed Networks*, SIAM INT'L CONF. ON DATA MINING 594, 594-602 (May 2019).

[23] Xin Wang et al., *Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory*, 2 INT'L JOINT CONF. ON NAT. LANGUAGE PROCESSING 1343, 1343–1353 (2015).

enable a potentially deeper understanding of factors that drive tax policy, something that can currently only be accomplished by Treasury and IRS employees.

For many years, researchers in the public domain have had access to the IRS Public Use File (PUF), which is based on a sub-sample of 1040 filers and is limited in feature representation due to risk of disclosing Personally Identifiable Information.[24] A fully synthetic data set for public use would overcome these limitations by constructing a population-level data set that preserves the underlying distributional properties of the actual data without risk to privacy through a disclosure violation. An effort is currently underway at the Urban Institute Tax Policy Center, in collaboration with IRS, to construct such a data set.[25] While significant computational challenges exist, this effort has the potential to accelerate tax research almost overnight.

d. Unsupervised Learning for Feature Discovery and Classification

This topic could easily be classified as a "grand challenge" for the IRS in the coming years. More data are being made available for analytics but with fewer labels for classification. Part of this phenomenon is due to an increasing amount of data collected, both structured and unstructured, with no inherent class labels. The IRS is also auditing fewer tax returns outside of the automated underreporter program, which matches returns with third-party information.[26] The remaining field audits, some of which are identified through machine learning techniques using historical data with class labels, are typically the

---

[24] *SOI Tax Stats - Individual Public-Use Microdata Files*, IRS (June 19, 2019), https://www.irs.gov/statistics/soi-tax-stats-individual-public-use-microdata-files.

[25] LEONARD E. BURMAN ET AL., TAX POLICY CTR., SAFELY EXPANDING RESEARCH ACCESS TO ADMINISTRATIVE TAX DATA: CREATING A SYNTHETIC PUBLIC USE FILE AND A VALIDATION SERVER (2018), https://www.urban.org/sites/default/files/publication/99247/safely_expanding_research_access_to_administrative_tax_data_creating_a_synthetic_public_use_file_and_a_validation_server_0.pdf.

[26] Robert A. Weinberger, *The IRS Data Book Tells a Story of Shrinking Staff, Fewer Audits, and Less Customer Service*, TAX POLICY CTR.: TAXVOX (June 7, 2019), https://www.taxpolicycenter.org/taxvox/irs-data-book-tells-story-shrinking-staff-fewer-audits-and-less-customer-service [https://perma.cc/DLU4-PGKD]

most complex but also provide a feature-rich source of data for supervised learning. For programs like Examination with decreasing amounts of labeled data, new approaches are needed.

Methodologically, there are numerous opportunities for unsupervised learning beyond classical methods of PCA, cluster analysis, mixture models, and density estimation. Recent advances in algebraic geometry, high-dimensional statistics, and convex optimization have produced a range of new techniques for nonlinear manifold learning that estimate subspace mappings from sparse, high dimensional data and overcome some of the well-known limitations of classical PCA.[27] Examples include kernel PCA, local linear embedding, spectral embedding (using the Laplacian of a graph), Hessian eigenmaps, t-distributed stochastic neighborhood embedding, and deep autoencoders. Beyond their use in dimension reduction, these techniques could be used for subspace anomaly detection (e.g., identity theft, misreporting on tax returns), where they have found practical success.[28]

Other considerations should include co-clustering via matrix factorization, subspace clustering,[29] frequent substructure discovery, and graph summarization (e.g., motifs and focused clustering of large attributed graphs).[30] For tax returns with high dimension, say over 2,000 line items, subspace clustering could be used to isolate important groupings based on subsets of features (e.g., of size 20, 30, 50, and so on). Frequent substructure discovery could be used to identify recurring graph patterns in a larger graph for applications in pattern matching, clustering, and anomaly detection. New algorithms, greater

---

[27] ALAN JULIAN IZENMAN, MODERN MULTIVARIATE STATISTICAL TECHNIQUES: REGRESSION, CLASSIFICATION, AND MANIFOLD LEARNING 598 (George Casella, Stephen Fienberg & Ingram Olkin, eds., Springer 1st ed., 2008); RENÉ VIDAL, YI MA & S. SHANKAR SASTRY, GENERALIZED PRINCIPAL COMPONENT ANALYSIS 171-345 (S.S. Antman, L. Greengard & P. Holmes eds., Interdisciplinary Applied Mathematics Ser. No. 40, 2016).

[28] David Charte et al., *A Practical Tutorial on Autoencoders for Nonlinear Feature Fusion: Taxonomy, Models, Software and Guidelines*, 44 INFO. FUSION 78, 78-96 (2018).

[29] T. Gayathri & D. LalithaBhaskari, *A Comprehensive Review of Subspace Clustering in the Analysis of Big Data*, 39 INT'L J. ENGINEERING TRENDS & TECH. 135, 135-142 (2016).

[30] Yike Liu, Tara Safavi, Abhilash Dighe & Danai Koutra, *Graph Summarization Methods and Applications: A Survey*, 51 ACM COMPUTING SURV. 44:1, 62:1-34 (2018).

use of distributed computing architectures, and GPU-based computing have made this notoriously difficult problem[31] more computationally tractable.[32]

The class of generative models that includes generative adversarial networks and variational autoencoders could be used to learn the distribution of an underlying set of data and provide additional opportunities for anomaly detection.[33] GANs, for example, have been adopted in financial, medical, cybersecurity, and other domains and have achieved success in producing convincingly real data that can be used in a semi-supervised to mimic minority class observations, thereby augmenting training sets for classification. Recent progress suggests their potential effectiveness for detecting threat patterns in ID theft, a serious problem that continues to plague both IRS and taxpayers.

e. A Field Guide for Analytics and Artificial Intelligence

In order to effectively exploit the vast amount of data, tools, and computational capabilities now available for decision making in tax administration, it helps for practitioners to know something about the full range of analytical techniques outside of traditional statistics and where they can be applied. Analytical literacy is as important for business analysts and research economists[34] as it is for agency leaders and is separate and distinct from the challenge of acquiring specialized skills. To be effective at creating a common vocabulary and providing exposure to a broad analytical tool shelf, a good field guide must not just be a context-free taxonomy of analytical techniques and should

---

[31] Frequent substructure discovery requires solving the subgraph isomorphism problem, which is NP-complete.

[32] Xuanhua Shi et al., *Graph Processing on GPUs: A Survey*, 50 ACM COMPUTING SURV. 81, 81:1-81:35 (2018).

[33] IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 651-716 (MIT Press et al. eds., 2016); Vít Škvára, Tomáš Pevný & Václav Šmídl, Are Generative Deep Models for Novelty Detection Truly Better? (July 13, 2018) (unpublished manuscript), https://arxiv.org/pdf/1807.05027.pdf.

[34] Susan Athey & Guido W. Imbens, *Machine Learning Methods Economists Should Know About*, 11 ANNU. REV. ECON. 685, 686-88, 709, 718, 720 (2019).

also include domain-specific examples, highlight pros and cons, and reference alternatives. Sentiment analysis belongs to natural language processing, but where can it be applied? Agent-based modeling is a simulation technique that can describe emergent behavior. What are the practical benefits?

f. Hybrid and Multi-stage Models

Ensemble learning has become a sophisticated and powerful approach for feature discovery and classification, with perhaps the best-known examples being boosting, bagging, and random forests for decision trees.[35] Beyond these blended approaches, multi-stage models also play an important role in the search for optimal classifiers and have gained increasing attention with deep learning. Classical examples include eigenvalue clustering from first-stage PCA as input to a supervised classifier, or singular value decomposition as a first-stage dimension reduction technique for topic modeling. Frequent substructure discovery has also been evaluated for this purpose.[36] During the Netflix Prize competition, a new standard emerged for recommender systems based on sparse matrix factorization as first-stage input to one or more second-stage optimization algorithms, a framework currently used by the IRS.[37] Recommender systems are also being used on graphs, as are word embeddings for feature engineering.[38] More work is needed to evaluate and compare advantages from hybrid and multi-stage models to provide useful insight to practitioners about where they can be most effective.

---

[35] CHRISTOPHER M. BISHOP, PATTERN RECOGNITION AND MACHINE LEARNING 653-4 (Michael Jordan et al. eds., 2006); TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION 624 (2nd ed. 2009).

[36] *See* Pang-Ning Tan, Hannah Blau, Steve Harp & Robert Goldman, *Textual Data Mining of Service Center Call Records*, 6 ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 417 (2000).

[37] Robert M. Bell, Yehuda Koren & Chris Volinsky, *The BellKor Solution to the NetFlix Prize*, AT&T LABS RES. (2007), https://www.netflixprize.com/assets/ProgressPrize2007_KorBell.pdf.

[38] Bryan Perozzi, Rami Al-Rfou & Steven Skiena, *DeepWalk: Online Learning of Social Representations*, 20 ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 701, 703, 709 (2014).

VI. BUILDING TALENT POOLS FOR THE NEXT GENERATION

The current quality of analytics in tax administration research at the IRS has room for improvement. There is inadequate consensus on the most important questions that can be addressed by advanced analytics and artificial intelligence, and insufficient application of known advanced techniques even in organizations with a clearly defined analytical mission. For tax administration to advance by maximizing a full range of data and analytical methods, greater investment in analytical skills is needed.

Creating a comprehensive talent strategy for analytics should consider the changing competitive landscape for skills in traditional curricula like statistics and computer science, as well as interdisciplinary needs between curricula, in the context of domain-specific problems. An effective strategy should include:

- **Competency framework**. Efforts are needed to build a workforce equipped with next- generation skills in advanced analytics and artificial intelligence. Candidates must be able to work with large-scale, heterogeneous, and multi-structured data; large memory and distributed computing architectures; and a range of modern analytical techniques beyond basic statistics. Such a framework should emphasize an interdisciplinary approach that includes domain knowledge and communication skills.

- **Job analysis and classification**. Current federal standards for job classification are inadequate to meet the increasing demand for deep analytical talent. Efforts are needed to determine core educational requirements of a next-generation workforce prepared to deal with new challenges that originate from multiple domains, and that can benefit from advanced analytics and artificial intelligence.

- **Qualified leadership**. It is an enormous challenge to find top leadership in today's competitive landscape of analytical talent.

To keep tax administration research on the cutting edge, priority should be given to recruiting leaders with a range of expertise in statistics, machine learning, modern computing architectures and software, and who demonstrate deep knowledge in the application of advanced analytics and artificial intelligence in other fields.

- **Industry and academic partnerships**. Progress has been made in recent years to deepen and expand collaborative research with industry and academia, catalyzing the application of novel and non-traditional approaches to tax administration and bringing needed enthusiasm for new scientific discovery. Yet, the amount of funding on such partnerships is a tiny fraction of what is required to create breakthrough research capabilities. And while many analytical problems are both exciting and rewarding, it is unlikely that top talent from universities could be attracted without a major increase in funding.

- **Agency cross-training**. The typical IRS career professional outside of the research community requires no experience in analytics, even when their job involves the use of data. For most employees, spreadsheets and custom IT applications are the most frequently used tools for decision making. While data scientists can learn the basics of a business process (e.g. notice delivery or Exam classification), the analytics training of an IRS accountant or business professional is not as easily accomplished. Efforts are needed to improve opportunities for temporary assignments, workshops, and short training courses in analytics.

- **Fellowship programs**. To attract critically needed talent in advanced analytics, consideration should be given to a two-year fellowship program, sponsored by the IRS Commissioner, with a salary commensurate with skills and experience. Candidates would need to hold a graduate degree in a suitable

field to qualify. The Food and Drug Administration has a good example of what such a program could look like.[39]

- **Boot camps for managers**. Efforts are needed to ensure that managers have access to training boot camps, short courses, and professional development workshops to learn the relevance of data, technology, and advanced analytical methods needed to shape business strategy.

[39] For more information, visit www.fda.gov.